



A deep learning algorithm for white matter hyperintensity lesion detection and segmentation

Yajing Zhang¹ · Yunyun Duan² · Xiaoyang Wang¹ · Zhizheng Zhuo² · Sven Haller^{3,4} · Frederik Barkhof^{5,6} · Yaou Liu²

Received: 14 June 2021 / Accepted: 24 September 2021 / Published online: 2 October 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Purpose White matter hyperintensity (WMHI) lesions on MR images are an important indication of various types of brain diseases that involve inflammation and blood vessel abnormalities. Automated quantification of the WMHI can be valuable for the clinical management of patients, but existing automated software is often developed for a single type of disease and may not be applicable for clinical scans with thick slices and different scanning protocols. The purpose of the study is to develop and validate an algorithm for automatic quantification of white matter hyperintensity suitable for heterogeneous MRI data with different disease types.

Methods We developed and evaluated “DeepWML”, a deep learning method for fully automated white matter lesion (WML) segmentation of multicentre FLAIR images. We used MRI from 507 patients, including three distinct white matter diseases, obtained in 9 centres, with a wide range of scanners and acquisition protocols. The automated delineation tool was evaluated through quantitative parameters of Dice similarity, sensitivity and precision compared to manual delineation (gold standard).

Results The overall median Dice similarity coefficient was 0.78 (range 0.64~0.86) across the three disease types and multiple centres. The median sensitivity and precision were 0.84 (range 0.67~0.94) and 0.81 (range 0.64~0.92), respectively. The tool’s performance increased with larger lesion volumes.

Conclusion DeepWML was successfully applied to a wide spectrum of MRI data in the three white matter disease types, which has the potential to improve the practical workflow of white matter lesion delineation.

Keywords White matter hyperintensity · Automated detection and segmentation · Multiple sclerosis · Multicentre · FLAIR

Yajing Zhang and Yunyun Duan contributed equally to the manuscript

✉ Yaou Liu
liuyaou@bjtth.org

¹ MR Clinical Science, Philips Healthcare, 258 Zhongyuan Rd, Suzhou, SIP, China

² Department of Radiology, Beijing Tiantan Hospital, Fengtai District, Capital Medical University, No. 119 the West Southern 4th Ring Road, Beijing 100070, China

³ Department of Imaging and Medical Informatics, University of Geneva, Geneva, Switzerland

⁴ Faculty of Medicine, University of Geneva, Geneva, Switzerland

⁵ Department of Radiology and Nuclear Medicine, Amsterdam University Medical Center, Amsterdam, The Netherlands

⁶ Queen Square Institute of Neurology and Center for Medical Image Computing, University College, London, UK

Introduction

MRI is widely used to detect white matter hyperintensity (WMHI) lesions in neurological disorders such as multiple sclerosis (MS) [1, 2], neuromyelitis optica spectrum disorders (NMOSD) [3, 4] and cerebral small vessel disease (CSVD) [5, 6]. Accurate identification of WMHI has clinical relevance for diagnosis and predicting prognosis, especially in early disease phases [7, 8].

In the current clinical workflow, although it is not challenging to manually identify white matter (WM) lesions, the large number of images as a daily workload is likely to cause decreased efficiency for radiologists. In addition, the delineation of WM lesions usually relies on manual labelling by experienced radiologists, which is time-consuming. Intra- and interrater reproducibility can be compromised by subjective judgement, the associated workload and the different experiences of raters [9, 10]. Various conventional machine

learning and deep learning methods have been proposed to automatically detect and segment WM lesions. Dadar et al. evaluated the segmentation performance from ten conventional linear and nonlinear classification techniques (naïve Bayes, logistic regression, decision trees, random forests, support vector machines, k-nearest neighbours, bagging and boosting) and observed superior performance from the random forest algorithm [11]. Recently, a few studies reported WMHI segmentation results using convolutional neural networks [12–16]. Rachmadi et al. proposed a 2-dimensional convolutional neural network (2D-CNN) scheme to segment WMHIs with no or mild vascular pathology and compared it with 15 other types of machine learning segmentation methods [17]. The 3D-UNet scheme has also been widely used for automated WM lesion segmentations, for example, in MS patient studies with T2-FLAIR and MP2RAGE images [18]. As the deep learning technique increasingly engendered such applications, a recent article with CLAIM guidelines (checklist for artificial intelligence in medical imaging) has been provided for such studies [19].

To be clinically useful, the segmentation technique needs to provide accurate results under heterogeneous imaging protocols for routinely available FLAIR images (e.g., 2D images with different slice thicknesses), which often calls for multicentre studies [20, 21]. Thus, there is a growing need to develop a fully automated tool that can deal with heterogeneous clinical data across various WM diseases. Unfortunately, previous studies have rarely demonstrated FLAIR-based deep learning WMHI lesion segmentation validated on multiple disease types and multicentre data.

The purpose of the study is to establish a deep learning (DL)-based automated WM lesion segmentation tool using a single-modality clinical FLAIR image that can be applied to data with WMHI lesions from multiple WM disease types (MS, NMOSD and CSVD) and from different clinical centres, with a wide range of vendors, image resolutions and scanning slice thicknesses. The tool aims to assist WM lesion detection and segmentation work for radiologists and clinicians.

Methods

Multicentre dataset

The MRI data from three disease types (MS, NMOSD and CSVD) that included WM lesions were collected from 9 centres (5 centres for MS and NMOSD and 4 centres for CSVD) (Table 1). MS diagnosis was determined according to the 2017 McDonald criteria [22]. The NMOSD diagnosis was based on the 2015 International Panel on NMOSD Diagnosis [23], and all patients had antibodies against AQP4

using the CBA method. CSVD diagnosis was based on the presence of white matter hyperintensity or more CSVD signs on MRI [24]. Patients who had other abnormalities, such as brain tumours on MRI and well-defined macrovascular stenosis on MRA, were excluded.

The current study data are from prospective studies, and the data partition was at the image level. The consent forms were signed by patients, and the data were anonymized. Routine clinical FLAIR images were acquired in axial orientation on 3.0 T or 1.5 T scanners from multiple vendors (Philips Achieva, GE Discovery MR750/Signa HDxt/Optima MR360 and Siemens Skyra/TrioTim). Each patient's FLAIR contained 17–30 slices to cover the whole brain. After data acquisition, one patient with MS was excluded due to poor image quality. One patient with NMOSD was excluded due to a history of brain trauma. Nine patients with CSVD were excluded due to > 50% intracranial macrovascular stenosis on MRA, and one CSVD patient was excluded due to incidental meningioma (Fig. S1). The final dataset consisting of 507 patients with 10,753 image slices took part in this multicentre study, including 135 MS patients (82 women; mean age (SD) 37.2 (12.7) years), 74 NMOSD patients (62 women; 39.5 (13.5) years) and 298 CSVD patients (177 women; 43.0 (16.1) years). The original data had a wide range of in-plane resolutions from 0.4102×0.4102 to 0.9375×0.9375 mm²/pixel, with slice thicknesses ranging from 3 to 8 mm. Details of the data parameters and distribution are shown in Table 1. As a preprocessing step, all FLAIR image slices were intensity-normalized and resampled to a matrix of 256×256 before network training.

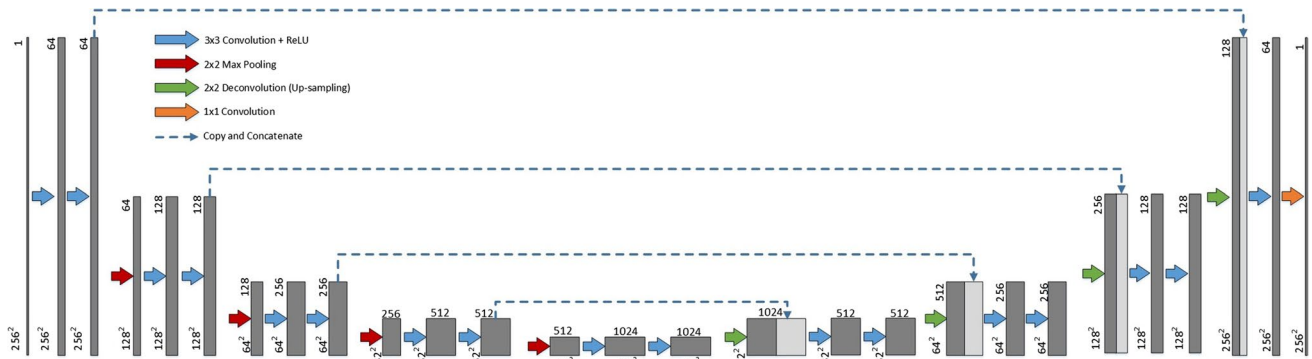
Manual labelling of the WM lesions was performed using 3D Slicer software [25]. Two experienced radiologists (Y. D with 12 years of experience and W. G with 10 years of experience) first completed a training session to reach a consensus on evaluating the imaging findings. After training, they performed manual labelling independently, and then the Dice index of the lesion masks from the two radiologists was calculated. The labels with relatively poor consistency (Dice < 0.85) need to be relabelled to reach a good consensus (Dice \geq 0.85) as the ground truth for further analysis.

Deep learning neural network

We employed a 2-D UNet strategy [26] with the network architecture shown in Fig. 1. It consisted of a contracting path and an expanding path with skip connections between them. The operation block in the contracting path consisted of two blocks of 3×3 convolutions with a rectified linear unit (ReLU) (blue arrows in Fig. 1). A 2×2 max pooling (red arrow in Fig. 1) was performed at the end of each block to downsample feature maps. In the expanding

Table 1 Data distribution from multiple clinical centres. 2D FLAIR images of MS, NMOSD and CSVD are collected on scanners from multiple vendors (Philips, GE, Siemens), with slice thickness rangingfrom 3 to 8 mm and in-plane resolution ranging from 0.4102×0.4102 to 0.9375×0.9375 mm²/pixel

Lesion type	Centre code	# patients	# image slices	Slice thickness (mm)	In-plane resolution (mm*mm/pixel)	Scanner type
MS	Age (mean (SD)) 37.2 (12.7) y	Gender—female ratio 82/135 (61%)				
	Ctr_MS#1	30	588	5.5, 6, 6.5	(0.4688*0.4688), (0.9375*0.9375)	GE Discovery MR750 3.0 T
	Ctr_MS#2	37	592	8	(0.4688*0.4688)	GE Discovery MR750 3.0 T
	Ctr_MS#3	27	541	6.5	(0.4492*0.4492), (0.7188*0.7188)	Siemens Skyra 3.0 T
	Ctr_MS#4	27	603	4, 5.2, 5.36, 5.5, 6.5	(0.4102*0.4102), (0.4297*0.4297), (0.4395*0.4395), (0.6875*0.6875)	Siemens Skyra 3.0 T
NMOSD	Age (mean (SD)) 39.5 (13.5) ys	Gender—female ratio 62/74 (84%)				
	Ctr_NMOSD#1	12	246	5, 6.5	(0.4688*0.4688), (0.5078*0.5078)	GE Discovery MR750 3.0 T
	Ctr_NMOSD#2	12	192	8	(0.4688*0.4688)	GE Discovery MR750 3.0 T
	Ctr_NMOSD#3	6	120	6.5	(0.7188*0.7188)	Siemens Skyra 3.0 T
	Ctr_NMOSD#4	43	1,373	4, 5, 6, 6.5, 7.5	(0.4297*0.4297), (0.4688*0.4688), (0.5*0.5), (0.8*0.8), (0.9375*0.9375)	Siemens TrioTim 3.0 T
CSVD	Age (mean (SD)) 43.0 (16.1) ys	Gender—female ratio 177/298 (59%)				
	Ctr_CSVD#1	52	1,067	6, 6.5	(0.4688*0.4688), (0.8976*0.8976)	GE Discovery MR750 3.0 T
	Ctr_CSVD#2	34	620	6.5, 7	(0.4492*0.4492), (0.4688*0.4688)	Philips Achieva 1.5 T
	Ctr_CSVD#3	70	1,562	6	(0.4492*0.4492)	Philips Achieva 1.5 T
	Ctr_CSVD#4	142	2,954	6.5, 7, 7.8, 8	(0.4688*0.4688)	GE Optima MR360/ Signa HDxt 1.5 T

**Fig. 1** The fully convoluted network architecture with U-net strategy. Each grey box represents a multi-channel feature map. Numbers on top of the boxes denote number of channels; width and height dimen-

sions are denoted at bottom. Light grey boxes represent the copied features through skip connections. The operations are denoted using different types of arrows

path, each operation block started with a 2×2 deconvolution for upsampling (green arrow in Fig. 1), followed by two blocks of 3×3 convolutions with ReLU. Skip connections (dashed line in Fig. 1) carried the features from the contracting path to the expanding path. The final layer was a fully convolutional layer with a 1×1 kernel (orange arrow in Fig. 1), which translated 64-channel features to a single-channel feature map. The output logits were compressed to the range 0–1 to predict the final lesion activation map.

Network training and testing

The FLAIR images from all the centres were used for training and testing sessions. In the training and validation steps, 60% and 20% of the image dataset were randomly selected from each centre and each disease type, respectively. This included data samples from all centres and all three disease types. This led to a total of 8,640 image slices included in the training and validation procedure. The remaining 20% of the dataset was used as the test set. UNet was trained with a loss function of combined binary cross-entropy loss and Dice coefficient loss [27]. The training was performed for 200 epochs using the Adam optimizer [28] with a starting learning rate of e^{-4} . Data augmentation was performed, including random horizontal flip and rotation (-10 to 10°). The model was implemented using the Python PyTorch 1.4 framework [29] with an NVIDIA GTX 1080 Ti*2 processor.

Performance evaluation

To quantitatively evaluate network performance, we calculated three evaluation metrics—the Dice similarity coefficient (DSC), the true positive rate (i.e. sensitivity) and the precision—on each image in the test set. The median of the three metrics was calculated along each disease type and each clinical centre.

Moreover, we partitioned the test set by different lesion volumes. For each 2D image, the lesion volume (LV) was calculated as the manually labelled lesion area multiplied by its slice thickness. LV groups were partitioned as follows: (G1) $LV < 0.2$ ml, (G2) $0.2 \text{ ml} \leq LV < 0.7$ ml, (G3) $0.7 \text{ ml} \leq LV < 2$ ml, (G4) $2 \text{ ml} \leq LV < 5$ ml and (G5) $LV \geq 5$ ml. The three metrics within each LV group were compared.

Statistical analysis

To explore the performance of our tool in the multicentre multidisease dataset, statistical analysis was applied to the three evaluation metrics using the Python SciPy package (<https://www.scipy.org>). We employed the Shapiro–Wilk test to check data normality and the Kruskal–Wallis test to evaluate the performance difference across disease types and groups based on lesion volume ranges. The Mann–Whitney tests were further performed as pairwise comparisons for those tests with significance (p value < 0.05), and Bonferroni correction was performed.

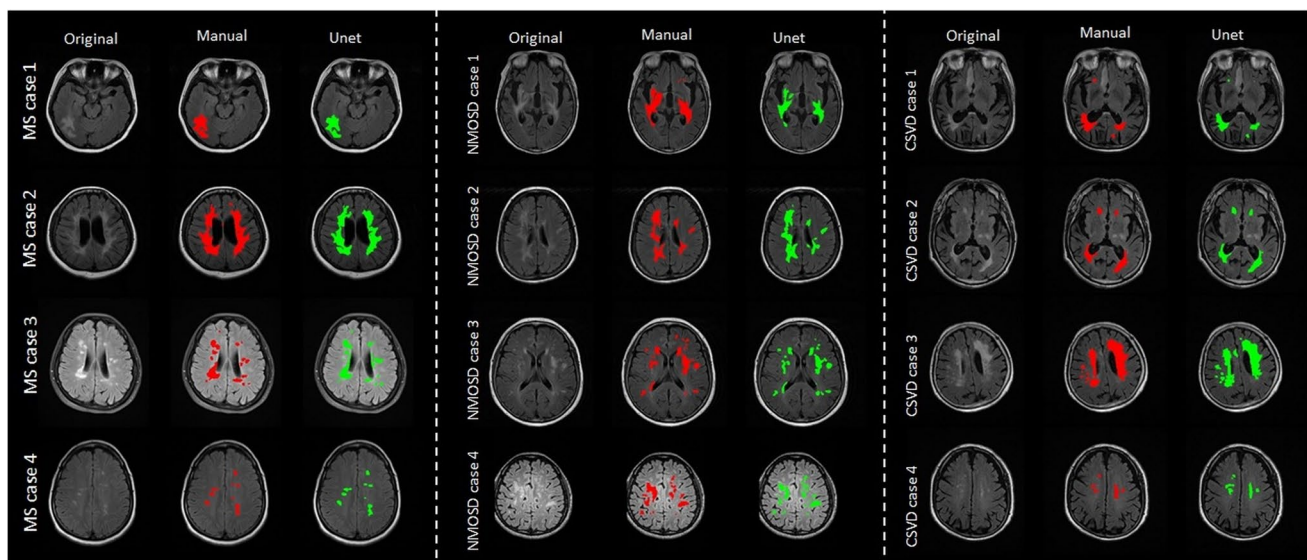


Fig. 2 Representative cases of WM lesion automated segmentation result (in green) vs. manual labelling (in red). Four typical cases in each disease type are randomly selected and covered both lower and upper position of the brain structures. The U-net-based automated

segmentation accords with the manual labels for the various lesion patterns and lesion sizes. The DSC for these cases are above 0.8. The segmentation time for each case is within 0.3 s

Table 2 Median DSC, sensitivity and precision of the WM lesions from three disease types of multicentre data. For DSC and sensitivity, there was no significant difference across the three disease lesion types ($p > 0.05$)

	MS										NMOSD					CSVD					Overall
	MS					NMOSD					CSVD										
	MS#1	MS#2	MS#3	MS#4	MS#5	NMOSD overall	NMOSD#1	NMOSD#2	NMOSD#3	NMOSD#4	CSVD overall	CSVD#1	CSVD#2	CSVD#3	CSVD#4						
Dice	0.80	0.79	0.79	0.75	0.88	0.86	0.77	0.80	0.80	0.83	0.75	0.78	0.79	0.74	0.78	0.78					
Sensitivity	0.83	0.73	0.82	0.80	0.87	0.92	0.84	0.79	0.78	0.86	0.86	0.85	0.86	0.81	0.86	0.84					
Precision	0.86	0.89	0.85	0.75	0.89	0.83	0.79	0.89	0.88	0.77	0.75	0.79	0.84	0.79	0.80	0.81					

Results

Figure 2 shows representative cases with WM lesions from MS (Fig. 2, left panel), NMOSD (Fig. 2, middle panel) and CSVD (Fig. 2, right panel), with manual labelling (overlay in red) and automated segmentations (overlay in green) for comparison. Four typical cases of each disease type were randomly selected. Our tool automatically segmented WM lesions in a manner consistent with manual labels, regardless of the variation in lesion patterns, locations, disease types and imaging parameters. The processing time for automated segmentation on each slice was within 0.3 s. For patient data with 17–30 slices in our dataset, the total time for whole-brain lesion segmentation was approximately 5–9 s.

Table 2 lists the quantitative results for each imaging centre. The median DSC for the testing dataset had an average value of 0.78, with 0.80 for MS lesions, 0.77 for NMOSD lesions and 0.78 for CSVD lesions. The median sensitivity and precision were 0.84 and 0.81, respectively. Figure 3 further details the distributions of the three performance metrics for each imaging centre (Fig. 3 A–C) and for each disease (Fig. 3 D–F). The Kruskal–Wallis test showed that there was no significant difference in DSC (p value = 0.09) (Fig. 3D) or sensitivity (p value = 0.54) (Fig. 3E) among the three disease types. However, the segmentation tool behaved differently among disease types in precision (p value = 0.009); in particular, the Mann–Whitney test demonstrated that the segmentation precision for MS lesions performed significantly better than for NMOSD and CSVD (p value after Bonferroni correction < 0.05) (Fig. 3F).

Figure 4 shows the segmentation performance for different WM lesion volumes, including all three disease types of lesions. The Kruskal–Wallis test showed a significantly higher performance (DSC, sensitivity and precision) as the lesion volume increased (p value < 0.001 for DSC; p value = 0.03 for sensitivity; p value < 0.001 for precision). As indicated in Fig. 4A (DSC), significant DSC improvements were found among the LV groups of G3 versus G1 ($p = 0.00025$), G4 versus G3 ($p = 2.4e^{-5}$) and G5 versus G4 ($p = 6.5e^{-6}$). For sensitivity (Fig. 4B), there was a significant improvement in G5 versus G4 ($p = 0.0024$). For precision, there was a significant improvement of G3 versus G1 ($p = 0.0002$) and G4 versus G3 ($p = 0.00011$).

Discussion

We presented DeepWML, an automated WM lesion delineation tool using a DL network that is largely agnostic to disease and scanners. The network was trained based on a large quantity of data with a wide range of imaging conditions, disease types and lesion sizes and therefore can be applied

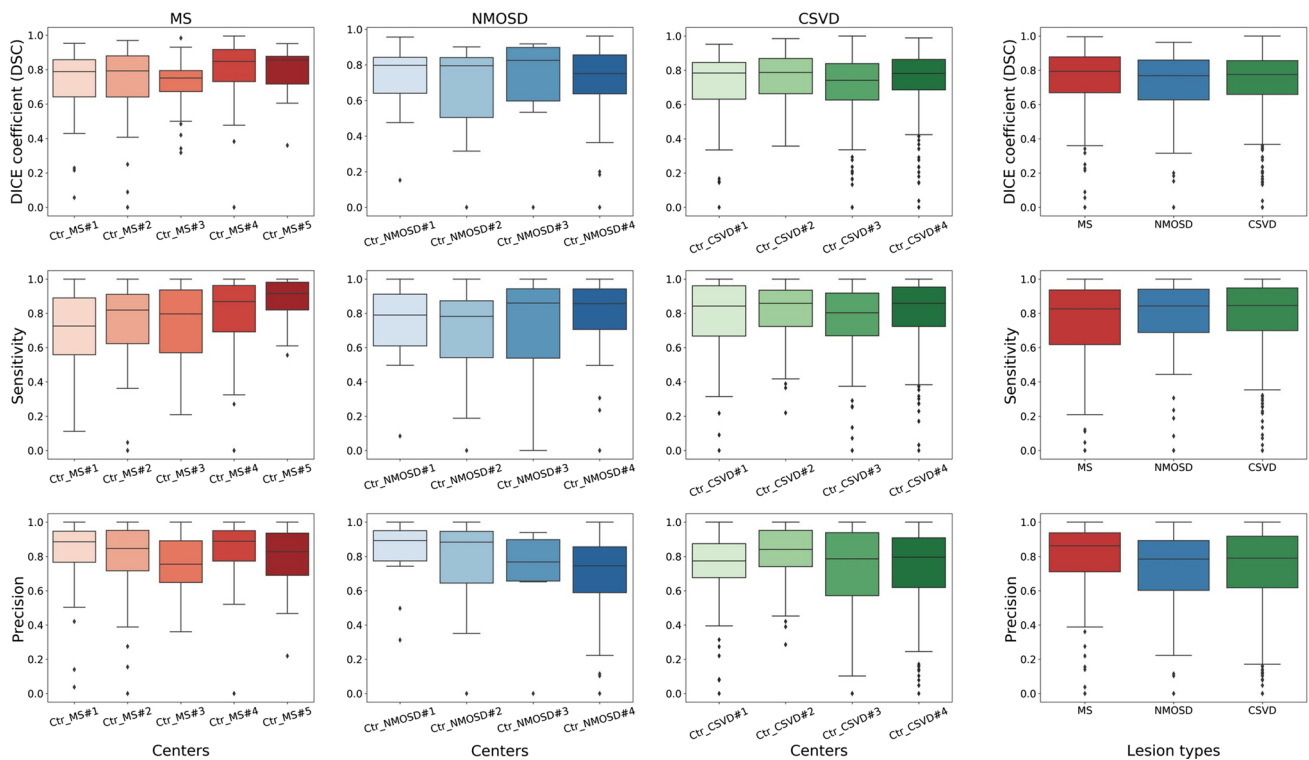


Fig. 3 Distribution of the DSC (top row), sensitivity (mid row) and precision (bottom row) across three disease types on multicentre data. Each column corresponds to centre-wise result for one type of disease. The right-most column shows result for three diseases. In

Fig. 3D and E, no significant difference in segmentation performance found in DSC or sensitivity ($p > 0.05$); in Fig. 3F, significantly better precision in MS lesion type is found ($*p < 0.05$ after Bonferroni correction)

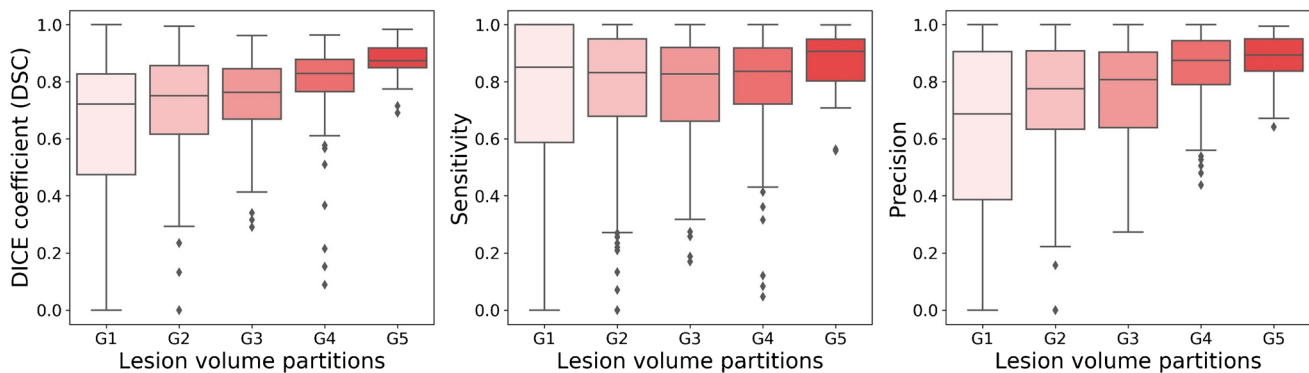


Fig. 4 DSC, sensitivity and precision with regard to lesion volume (LV) partitions. The testing data were partitioned according to the manually labelled WM lesion volumes with 5 groups: (G1) $LV < 0.2$ ml; (G2) $0.2 \text{ ml} \leq LV < 0.7$ ml; (G3) $0.7 \text{ ml} \leq LV < 2$ ml;

(G4) $2 \text{ ml} \leq LV < 5$ ml; (G5) $LV \geq 5$ ml. The segmentation performance improves along with the increasing WM lesion volume groups, with detailed statistical results marked in the figures. $*p < 0.05$ after Bonferroni correction

to a wide spectrum of MRI data for automated WM lesion segmentation in three main WM diseases (MS, NMOSD and CSVD).

In this study, we employed three evaluation metrics (DSC, sensitivity and precision) to quantify the segmentation

performance. The DSC depicts the overlap between the manual labelling and the automated result, and the median DSC reached 0.78 for the overall multicentre dataset. This is comparable to the results reported from alternative approaches [15].

Moreover, a good median sensitivity (84%) ensured that most of the lesion pixels were correctly identified; a good precision (81%) signified that there were few false-positive pixels. Importantly, the computational time per segmentation slice was within 0.3 s (5–9 s for each patient), which compares favourably with other DL algorithms. The few failure cases with discordant segmentation results against the manual labels (shown in Fig. S2) were analysed, and there were the following three causes of detection failure. First, some extremely small WM lesions that occupy less than 10 pixels (Fig. S2, panel A) are difficult to detect by the current DeepWML model, probably due to the small portion of training data with extremely small lesion sizes. Second, some WMHs were contaminated with normal tissue around the ventricle area (Fig. S2, panel B) because they share the high-intensity feature. The third type of failure case was due to the subtle intensity difference in WM lesions against their surrounding tissues (Fig. S2, panel C). For the performance with different lesion volumes, our study reached a median Dice of 0.87 for lesion volume > 5 ml and 0.71 for lesion volume < 0.2 ml, which is a good performance for such small lesions. Small lesions below the in-plane resolution are confusing for both the automatic tool and manual labelling. For the spatial accuracy expressed by the Dice coefficient, our study showed a trend of increased median Dice along with the lesion volume increase (Fig. 4, left). This result is highly consistent with the finding from previous work [30], in which “a clear trend for worse performance at lower volumes” was stated. A larger dataset is recommended to further confirm this finding in the subgroup analysis. Our algorithms showed several advantages to be applied in clinical practice. First, many previous studies used multiple MR image contrasts to enrich the input information to achieve better segmentation performance on MS patient data [16, 31, 32]. However, multicontrast MR data require additional scan time and usually need coregistration to integrate information from multiple channels. The input of our tool only requires routine 2D FLAIR images and can be applied for images from multiple scanners with different vendors, with a wide range of in-plane resolution, slice thickness and other imaging parameters. Second, the applicability of the trained UNet tool is not limited to WM lesions of one specific disease type. Our results showed no significant difference in the overlap ratio between automated and manual results (DSC) from MS, NMOSD and CSVD, implying that it is not required to have specified diagnostic information before commencing WM lesion segmentation. Last, the tool is not dependent on skull stripping the images, which makes preprocessing easier. The abovementioned features indicate great potential for our tool to be used in common clinical scenarios.

There are several limitations to our tool. First, the proposed network is based on 2D images, as well as the

evaluation of the lesions on the 2D slice level. It is possible that expansion to a 3D network may achieve similar or better performance. One reason that we adhere to the 2D network is that there is a large variation in the slice thickness in our routine clinical 2D FLAIR images (3 mm ~ 8 mm). This leads to a very limited number of slices for certain patients, and the 2D strategy may fit this situation. Second, the trained network was tested on the three common white matter disease types (MS, NMOSD and CSVD), while the performance of our tool in less common types of white matter diseases may need further validation. Last, the tool performance with small lesion volumes (e.g. less than 0.2 ml) needs to be improved.

Conclusion

In summary, we developed DeepWML, a DL-based automated WM lesion segmentation tool for WM lesion identification and delineation with satisfactory performance in three main WM diseases. The robustness of the tool and the computation efficiency would allow integration into the routine clinical workflow.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00234-021-02820-w>.

Acknowledgements Frederik Barkhof is supported by the NIHR biomedical research center at UCLH. The authors are grateful to Dr. Susumu Mori for his helpful suggestions.

Author contribution All authors contributed to the study conception and design. Data collection, methodology and data analysis were performed by Yajing Zhang, Yunyun Duan and Xiaoyang Wang. The first draft of the manuscript was written by Yajing Zhang and Yaou Liu, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by National Natural Science Foundation of China (grant numbers: 81571631, 81870958), Beijing Nova Program (grant number: xx2013045) and Natural Science Foundation of Beijing Municipality (grant number: 7133244).

Data availability Data that support the findings in this study are available from the corresponding author upon reasonable request.

Code availability The source code generated during the current study is available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare no competing interests.

Ethics approval This study was approved by the institutional review boards of Beijing Tiantan Hospital, Capital Medical University, Beijing, China. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Consent to participate Written informed consent was obtained from each participant.

Consent for publication Consent for publication was obtained from all participants.

References

- Rovira A, León A (2008) MR in the diagnosis and monitoring of multiple sclerosis: an overview. *Eur J Radiol* 67(3):409–414
- Wattjes MP, Rovira A, Miller D, Yousry TA, Sormani MP, de Stefano MP et al (2015) Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis—establishing disease prognosis and monitoring patients. *Nat Rev Neurol* 11(10):597–606
- Asgari N, Skejoe HPB, Lillevang ST, Steenstrup T, Stenager E, Kyvik KO (2013) Modifications of longitudinally extensive transverse myelitis and brainstem lesions in the course of neuromyelitis optica (NMO): a population-based, descriptive study. *BMC Neurol* 13:33–33
- Dutra BG, da Rocha AJ, Nunes RH, Maia ACMJ (2018) Neuromyelitis optica spectrum disorders: spectrum of MR imaging findings and their differential diagnosis. *Radiographics : a review publication of the Radiological Society of North America, Inc.* 38(1): p. 169–193
- Zhang X, Tang Y, Xie Y, Ding C, Xiao J, Jiang X et al (2017) Total magnetic resonance imaging burden of cerebral small-vessel disease is associated with post-stroke depression in patients with acute lacunar stroke. *Eur J Neurol* 24(2):374–380
- Li G, Zhu C, Li J, Wang X, Zhang Q, Zheng H et al (2018) Increased level of procalcitonin is associated with total MRI burden of cerebral small vessel disease in patients with ischemic stroke. *Neurosci Lett* 662:242–246
- McKinley R, Wepfer R, Grunder L, Aschwanden F, Fischer T, Friedli C et al (2019) Automatic detection of lesion load change in multiple sclerosis using convolutional neural networks with segmentation confidence. *NeuroImage Clinical* 25:102104–102104
- Salem M, Valverde S, Cabezas M, Pareto D, Oliver A, Salvi J et al (2019) A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. *NeuroImage Clinical* 25:102149–102149
- Zijdenbos AP, Forghani R, Evans AC (2002) Automatic “pipeline” analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. *IEEE Trans Med Imaging* 21(10):1280–1291
- García-Lorenzo D, Francis S, Narayanan S, Arnold DL, Collins DL (2013) Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Med Image Anal* 17(1):1–18
- Dadar M, Maranzano J, Misquitta K, Anor CJ, Fonov VS, Tartaglia MC, et al. Performance comparison of 10 different classification techniques in segmenting white matter hyperintensities in aging. (1095–9572 (Electronic)).
- Diniz PHB, Valente TLA, Diniz JOB, Silva AC, Gattass M, Ventura N et al (2018) Detection of white matter lesion regions in MRI using SLIC0 and convolutional neural network. *Comput Methods Programs Biomed* 167:49–63
- Li H, Jiang G, Zhang J, Wang R, Wang Z, Zheng W-S et al (2018) Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. *Neuroimage* 183:650–665
- Xu B, Chai Y, Galarza CM, Vu CQ, Tamrazi B, Gaonkar B et al (2018) Orchestral fully convolutional networks for small lesion segmentation in brain MRI. . *Proceedings IEEE International Symposium on Biomedical Imaging* 2018:889–892
- Duong MT, Rudie JD, Wang J, Xie L, Mohan S, Gee JC et al (2019) Convolutional neural network for automated FLAIR lesion segmentation on clinical brain MR imaging. *AJNR Am J Neuroradiol* 40(8):1282–1290
- Schmidt P, Gaser C, Arsic M, Buck D, Förchler A, Berthele A et al (2012) An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. *Neuroimage* 59(4):3774–3783
- Rachmadi MF, Valdés-Hernández MdC, Agan MLF, Di Perri C, Komura T (2018) Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain MRI with none or mild vascular pathology. *Computerized Medical Imaging and Graphics*. 66: p. 28–43
- La Rosa F, Abdulkadir A, Fartaria MJ, Rahmanzadeh R, P.-J. Lu, R. Galbusera, et al. (2020). Multiple sclerosis cortical and WM lesion segmentation at 3T MRI: a deep learning method based on FLAIR and MP2RAGE. *NeuroImage: Clinical*. 27: p. 102335
- Mongan JA-O, Moy LA-O, Kahn CEJA-O. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. (2638–6100 (Electronic))
- Heinen R, Steenwijk MD, Barkhof F, Biesbroek JM, van der Flier WM, Kuijff HJ et al (2019) Performance of five automated white matter hyperintensity segmentation methods in a multicenter dataset. *Sci Rep* 9(1):16742–16742
- Le M, Tang LYW, Hernández-Torres E, Jarrett M, Brosch T, Metz L et al (2019) FLAIR(2) improves LesionTOADS automatic segmentation of multiple sclerosis lesions in non-homogenized, multi-center, 2D clinical magnetic resonance images. *NeuroImage Clinical* 23:101918–101918
- Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T, Comi G et al (2018) Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology* 17(2):162–173
- Wingerchuk DM, Banwell B, Bennett JL, Cabre P, Carroll W, Chitnis T et al (2015) International consensus diagnostic criteria for neuromyelitis optica spectrum disorders. *Neurology* 85(2):177–189
- Wardlaw JM, Smith EE, Biessels GJ, Cordonnier C, Fazekas F, Frayne R et al (2013) Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology* 12(8):822–838
- Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S et al (2012) 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging* 30(9):1323–1341
- Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing
- Sudre CH, Li W, Vercauteren T, Ourselin S, Cardoso MJ (2017) Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham: Springer International Publishing
- Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. *arXiv e-prints*: p. arXiv:1412.6980
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv e-prints*: p. arXiv:1912.01703
- Ribaldi F, Altomare D, Jovicich J, Ferrari C, Picco A, Pizzini FB et al (2021) Accuracy and reproducibility of automated white matter hyperintensities segmentation with lesion segmentation tool: A European multi-site 3T study. *Magn Reson Imaging* 76:108–115
- Steenwijk MD, Pouwels PJW, Daams M, van Dalen JW, Caan MWA, Richard E et al (2013) Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). *NeuroImage Clinical* 3:462–469
- Damangir S, Westman E, Simmons A, Vrenken H, Wahlund L-O, Spulber G (2017) Reproducible segmentation of white matter hyperintensities using a new statistical definition. *Magma (New York, N.Y.)*. 30(3): p. 227–237

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.